Master's thesis

Multilingual model using cross-lingual word embeddings based on subword alignment and cross-task projection

（単語の表層アライメントとタスク横断写像に基づく多言語単語分散表現を用いた多言語モデル）

48-176447 Jin Sakuma

Department of Information and Communication Engineering
Graduate School of Information Science and Technology
The University of Tokyo

Supervisor
Associate Professor Naoki Yoshinaga

January 31, 2019

Abstract

While deep learning achieves high accuracy in various natural language tasks via representation learning with massive datasets, it is impractical to prepare such large-scale datasets for every pair of task and language. Multilingual models mitigate this problem by transferring models trained in resource-rich languages to resource-poor languages. The standard approach to obtaining multilingual models is to train a neural network with embedding layers fixed to language-independent word representations, namely, cross-lingual word embeddings.

In this study, we consider two challenges of this approach; cross-lingual word embeddings obtained via existing methods degrade for distant language pairs, and this method fails to exert true potential of neural networks as the embedding layer is not optimized for the given task. To mitigate these problems, we propose a unsupervised method to improve cross-lingual word embeddings for distant language pairs using subword alignment (§ 4) and a method to obtain task-specific cross-lingual word embeddings to obtain a fully task-specific multilingual model (§ 5).

In order to improve the quality of cross-lingual word embeddings for distant language pairs, we first capture unambiguously-translatable word pairs such as loanwords and named entities to constitute a more reliable bilingual dictionary to induce cross-lingual word embeddings. We then induce cross-lingual word embeddings from the obtained unambiguously-translatable bilingual dictionary. Experimental results in four language pairs, English-Japanese, English-Finnish, English-Spanish, and English-Italic, indicate that cross-lingual word embeddings obtained with our method were more accurate than those obtained by the state-of-the-art method, especially on distant language pairs.

To obtain a fully task-specific multilingual model, we use a cross-task projection that maps pre-trained cross-lingual word embeddings to the task-specific embedding layer of the neural network trained on the resource-rich language. Experimental results on sentiment analysis and document classification tasks demonstrated that our method obtained superior performance compared to traditional models with fixed embedding layers.

Combining these two methods, we obtain a multilingual model that is fully task-specific (including the embedding layer) and performs well for distant language pairs.

Keyword: natural language processing, multilingual models, cross-lingual word embeddings

Acknowledgements

I would like to use this part of the master's thesis to express my gratitude to everyone who helped me through my two years of master's program. First of all, I would like to show the greatest gratitude to my adviser, Professor Naoki Yoshinaga for precise advice and patiently guiding me. Every time I discussed with him, I gain some insights in this area and it helped me to develop my papers. I am grateful to Professor Masashi Toyoda for his support and advise and to Professor Masaru Kitsurekawa for the best environment for research.

I would also like to thank all members of the Kitsurekawa-Toyoda-Nemoto-Yoshinaga Lab. They helped me a lot in various aspects inside and outside of the lab. We often had great discussions which helped me a lot to organized my thoughts. Also, when my research did not go well, they cheered me up to give me the power to continue my work.

Finally, I want to thank my family and friends for their support. My friends were always there for me through tough times, and my family supported me significantly. I could not complete this thesis without any of these people.

January 31, 2019

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Multilingual model for cross-lingual resource utilization

Various tasks in natural language process (NLP) have experienced significant improvements via representation learning with deep learning on a massive annotated corpus. However, preparing such a corpus in every language for each task is impractical, and as a result, the performances of NLP models in many resource-poor languages are limited. A multilingual model, which can be trained in a resource-rich language (hereafter, source language) and then applied to another resource-poor language (hereafter, target language), mitigate this problem because it utilizes resources across languages.

Obtaining a multilingual model should require not only minimal annotated resources in the target language but also minimal cross-lingual resources such as bilingual dictionary and parallel corpus because obtaining cross-lingual resources also requires much human labor. In this study, we assume there is neither annotated resource in the target language nor cross-lingual resource across the source and target language to address the most common situation.

The most common method (especially in an unsupervised scenario) to obtain a multilingual model is to exploit cross-lingual word embeddings to absorb the difference in the vocabularies across languages. Cross-lingual word embeddings represent words from multiple languages as vectors such that words (no matter which languages they are from) are close to each other if they have a similar meaning. A neural network model trained while fixing the embedding layer to the pre-trained cross-lingual word embeddings can be applied in the language other than the training language.

This method of obtaining a multilingual model fails to perform well due to (1) the quality of cross-lingual word embeddings trained in an unsupervised manner perform poorly for distant language pairs and (2) it fails to induce task-specific word embeddings because we fix the embedding layer during the training. Recent studies indicate that cross-lingual word embeddings are obtainable in a fully unsupervised manner for similar language pairs, but for distant language pairs such as English-Japanese or English-Finnish, the quality degrades significantly [1, 2]. Also, previous studies [3, 4] and our experimental results indicate that a neural network with the embedding layer fixed to pre-trained word embeddings performs poorly. This study addresses these two problems by proposing two methods respectively, **namely cross-lingual word embeddings based on subword alignment** and **a fully task-specific multilingual model using a cross-task projection of cross-lingual word embeddings**.

1.2 Cross-lingual word embeddings based on subword alignment

Existing studies [1, 2] learn cross-lingual word embeddings from bilingual dictionary induced in an unsupervised manner. We believe that this method fails to be effective in distant language pairs because the bilingual dictionary induced for distant language pairs is noisy. The polysemous words in distant languages are likely to

share only a part of their senses, and the remaining senses are irrelevant to each other. For example, an English word "moon" has multiple translations in Japanese such as "月 (The moon)," and "衛星 (satellite)," while "月" has multiple translations in English such as *Monday* and *month*, which are not included in the meaning of "moon."

To mitigate this problem of ambiguous translations in distant language pairs, we take advantage of words that have surface correspondences such as loanwords and named entities and use these word pairs for inducing cross-lingual word embeddings. We assume that such word pairs with surface correspondences are likely to be unambiguously translatable with each other since those words originally come from the other language. To extract such words from the bilingual dictionary, we exploit subword alignment to extract *well-aligned* word pairs in the bilingual dictionary.

1.3 Fully task-specific multilingual model using cross-task projection of cross-lingual word embeddings

A multilingual model obtained in the method mentioned above fails to induce a task-specific representation of words as we must fix the embedding layer during the training in the source language. To exploit a fully task-specific neural network in cross-lingual settings, we propose a novel method of projecting pre-trained cross-lingual word embeddings to word embeddings of a task-specific neural network that is trained for the target task with the training data in a source language. We then utilize the obtained cross-task projection to obtain task-specific cross-lingual word embeddings of the target language that can be used for the task-specific neural network.

To obtain the above cross-task projection of cross-lingual word embeddings, we propose a simple, yet effective method of locally-linear mapping. This method is

built on the assumption that local topology is preserved between the semantic spaces of word embeddings in two NLP tasks; in other words, *adequately similar* words in pre-trained cross-lingual word embeddings will have similar representation even in task-specific semantic space.

1.4 The structure of this paper

The structure of this paper is as follows. In § 2, we present previous studies on cross-lingual word embeddings and multilingual models, and discuss the relationships with this work. In § 3, we introduce monolingual word embeddings and an unsupervised method to obtain cross-lingual word embeddings [1] that we adopted for this study. In § 4, we propose our method that utilizes subword alignment to obtain cross-lingual word embeddings in an unsupervised manner and then conduct experiments to evaluate this method and understand its characteristics. In § 5, we propose a method that induces task-specific cross-lingual word embeddings in order to obtain a fully task-specific multilingual model, and then conduct experiments to evaluated effect of the task-specific word representation in multilingual models. In § 6, we summarize our study and discuss our future work to further improve the quality of multilingual models.

Chapter 2

Related Work

In this chapter, we introduce previous studies that induce cross-lingual word embeddings (§ 2.1) and multilingual models (§ 2.2), and discuss their relationships to this study.

2.1 Cross-lingual word embeddings

Supervised methods Most of the methods to obtain cross-lingual word embeddings assume that some cross-lingual resources are available. These methods basically obtain monolingual word embeddings independently for each language, and then learn mappings across languages so that word pairs in a bilingual dictionary are close to each other after mapping.

The most simple way to learn such mappings is a regression method which maps embeddings of a language into another using least square objective [5–7]. Faruqui and Dyer extended this method to use canonical correlation analysis to map both languages into one semantic space [8], and Lu et al. proposed to use deep canonical correlation analysis to further enhance the mappings [9]. Recent studies [10, 11] follow the regression method but with a constraint of the mapping to be orthogonal

which not only improves the quality of resulting embeddings but also makes it possible to analytically compute the globally optimal mapping via singular value decomposition. Artetxe et al. generalized these methods as part of a multi-step framework [12].

The other stream of studies attempts to train word embeddings of multiple languages simultaneously instead of mapping independently induced monolingual word embeddings. MultiCluster first obtains multilingual corpus by assigning the same word ID for word pairs in the hand-crafted bilingual dictionary and then trains Skip-gram model on the multilingual corpus [13]. However, this method fails to perform well due to polysemies, and to mitigate this, Duong et al. dynamically replaced the target word with its translation during the training of Skip-gram model [14]. Another study utilizes earth mover's distance to enhance the quality of resulting cross-lingual word embeddings instead of replacing words [15]. Nakashole improves cross-lingual word embeddings for distant languages in a supervised situation by incorporating neighborhood information in the pre-trained embeddings [16].

Semisupervised methods Practically, it is very human-intensive to obtain cross-lingual resources, and thus some researches focus on obtaining cross-lingual word embeddings with minimal supervision. Artetxe et al. successfully obtain cross-lingual word embeddings from 25 words pairs or numerals by self-learning framework which alternatively induce dictionary and train mapping [17]. Another study utilized word pairs with exactly the same character string [18] which can be obtained automatically. However, this method is not applicable to language pairs with different symbolic systems such as English-Japanese, and the experimental results indicate that their performance is limited.

Unsupervised methods Unsupervised learning of cross-lingual word embeddings is first obtained using earth movers distance as the objective instead of least square objective [19]. In this method, Wasserstein GAN [20] was used to induce the mapping across languages which minmizes the difference between distributions of

the embeddings of the source and the target languages. Other studies exploit adversarial learning to obtain cross-lingual word embeddings without any cross-lingual supervision [2, 21] and Artetxe et al. enhanced self-learning framework with unsupervised initialization strategy and robust learning method [1]. Instead of learning a mapping between pre-trained monolingual word embeddings, Wada and Iwata train a language model sharing some of the parameters of the neural network and the resulting embedding layer is taken as cross-lingual word embeddings [22]. These unsupervised methods sometimes exhibited better performance against ones based on cross-lingual resources; in other words, cross-lingual resources are not always optimal for obtaining cross-lingual word embeddings.

While these unsupervised methods obtain high-quality cross-lingual word embeddings for similar language pairs (typically, European languages), performances are still limited in distant language pairs such as English-Japanese. In this work, we focused on such distant language pairs and improved cross-lingual word embeddings with subword alignment.

Task-specific word embeddings The effort to obtain task-specific cross-lingual word embeddings has been made previously. Gouws et al. obtain task-specific cross-lingual word embeddings by constructing a task-specific bilingual dictionary, which defines equivalent *classes* designed for the given task instead of equivalent *semantics* [23]. All words (regardless of their languages) in the same equivalence class get the same vector representations. For instance, equivalence classes for POS tagging task equate two words in a different language if they have overlapping syntactic categories. Although they successfully obtained task-specific cross-lingual word embeddings for POS tagging, it is not clear how we should define the task-specific bilingual dictionary for other NLP tasks, and preparing them is usually human-intensive.

2.2 Multilingual models

Lack of resources in many languages is a deeply rooted problem in natural language processing, and there have been many pieces of researches contributed to mitigating this problem by transferring models across languages.

Multilingual models using parallel corpus An intuitive approach to realize the cross-lingual transfer of a model is to utilize machine translation [24, 25]. They either translate annotated corpus of the source language to the target language and train a model in the translated corpus, or train a model in the source language and translate input in the prediction. Johnson et al. learned multilingual neural machine translation system from parallel corpus and enabled zero-shot translation; they can translation among language pairs where no parallel corpus is available [26]. Other studies utilize parallel corpus directly to train multilingual model instead of training machine translation systems [27, 28]. While some of these methods do not rely on an annotated corpus in the target language, they heavily rely on cross-lingual resources such as parallel corpus.

Multilingual models with cross-lingual word embeddings Another intuitive method to obtain multilingual models is to fix the embedding layer of a neural network to cross-lingual word embeddings. Many existing pieces of research implemented this for various tasks in an unsupervised senario [14] where no annotated corpus is available in the target language as ours and a supervised scenario [29, 30]. Other studies enhanced these models by employing language-adversarial networks [31, 32]

As discusses in § 5.2, these models fail to induce task-specific representation of words, and thus cannot exert true potential of neural networks.

Multilingual models with character embeddings Several studies utilize character level embeddings shared across languages to obtain multilingual models [33, 34]. An obvious weak point of these methods is that they do not apply to distant language pairs of ones with a different alphabet while our method only relies on cross-lingual word embeddings which are obtainable regardless of the alphabet of the language [1].

Chapter 3

Preliminaries

In this chapter, we introduce methods to obtain monolingual word embeddings from a raw corpus in § 3.1, and an unsupervised method to obtain cross-lingual word embeddings in § 3.2.

3.1 Monolingual word embeddings

Monolingual word embeddings are vector representation of words in a language so that similar semantic words are close to each other as vectors. Most of the methods to obtain monolingual word embeddings learn the representation from a raw corpus such as Wikipedia[1] and Twitter[2] by exploiting Distributional Hypothesis [35] which states that the semantics of words can be implied from the surrounding words. Here, we introduce Skip-gram [36] and its extension, Subword Information Skip-gram [37], which we adopted for this study.

[1] https://dumps.wikimedia.org/
[2] https://twitter.com/

3.1.1 Skip-gram

In Skip-gram model, each word w has two vectors X_w^{trg} and X_w^{ctx} and the model is trained to predict the surrounding (context) word from a target word in the raw corpus. Suppose that $D = \{w_1, w_2, \cdots, w_N\}$ are the list of words in the raw corpus and let w_i be the target word. We optimize the model to predict the context words $\{w_{i-k}, w_{i-k+1}, \cdots, w_{i-1}, w_{i+1}, \cdots, w_{i+k-1}, w_{i+k}\}$ from the target word w_i.

Typically, the negative sampling technique is exploited for computational efficiency. Given the target word, we randomly sample words (negative samples) and the model is trained to predict the word is a negative sample or an actual context word. We optimize X^{trg} and X^{ctx} by stochastic gradient descent to minimize the following objective function

$$L = \sum_{i=1}^{N} \sum_{k \in \{-K, \cdots, -1, 1, \cdots, K\}} \log P(t = 1 | w_i, w_{i+k}) + \sum_{j=0}^{NS} \log P(t = 0 | w_i, w'_{ij})$$

where NS is the number of negative samples. We compute the probability of the word pair v, c to be an actual neighboring word pair from the raw corpus by

$$P(t = 1 | v, c) = \sigma(X_v^{\mathrm{trg}} \cdot X_c^{\mathrm{ctx}})$$

where $\sigma(\cdot)$ is the sigmoid function, $\sigma(x) = \frac{1}{1+\exp(-x)}$. The probability of the word pair v, c not to be an actural neighboring word pair is

$$\begin{aligned} P(t = 0 | v, c) &= 1 - P(t = 1 | v, c) \\ &= 1 - \sigma(X_v^{\mathrm{trg}} \cdot X_c^{\mathrm{ctx}}) \\ &= \sigma(-X_v^{\mathrm{trg}} \cdot X_c^{\mathrm{ctx}}) \end{aligned}$$

After the optimization, we use X^{trg} as the word embeddings.

3.1.2 Subword-Information Skip-gram

Subword-Information Skip-gram [37] is an extension of Skip-gram by utilizing subword information. This method not only improves the quality of the word embeddings, but it also enables to induce word embeddings of unknown words.

This method optimizes the same objective function as Skip-gram (§ 3.1.1) but instead of words, character n-grams have embeddings. Word embeddings are computed as the summation of embeddings of its n-grams.

For each word in the raw corpus, special characters `<` and `>` are added to the beginning and end of the word, and then they compute the bag-of-character n-grams, $\mathcal{G}$. The word itself is also included in $\mathcal{G}$. For example, from the word *where* with $n = 3$, the resulting $\mathcal{G}$ would be

$$\mathcal{G} = \{\texttt{<wh}, \texttt{whe}, \texttt{her}, \texttt{ere}, \texttt{er>}, \texttt{<where>}\}.$$

The embedding X_s^{trg} of a word s is computed by

$$X_s^{\text{trg}} = \sum_{g \in \mathcal{G}} X_g^{\text{ngram}}$$

where X_g^{ngram} is the vector representation of n-gram g.

Given the trained embeddings, the embeddings of unknown word can be computed as the summation of embeddings of all its character n-grams.

3.2 Unsupervised learning of cross-lingual word embeddings

The role of cross-lingual word embeddings in a multilingual model is to represent words in two languages with fix-length vectors in a shared semantic space to absorb

the difference among vocabularies across languages. They represent semantically-similar words as vectors with similar values irrespective of languages of the words. In this study, we use cross-lingual word embeddings as a resource to obtain task-specific cross-lingual word embeddings.

In this study, we adopted the state-of-the-art method of obtaining cross-lingual word embeddings of two languages [1]. This method learns cross-lingual word embeddings from pre-trained word embeddings in both languages in an unsupervised manner without any cross-lingual resources such as bilingual dictionaries. The method can be applicable to our target scenario that assumes no language resource for the target (resource-poor) language.

First, word embeddings X and Y of the source and target language are obtained through existing methods of learning word embeddings via unsupervised tasks such as language modeling [36]. Then they learn linear orthogonal mappings W_x and W_y in an iterative manner so that the mapped embeddings XW_x^T and YW_y^T are in the same semantic space and semantically-similar words in different languages have similar vector representation. They estimate an initial bilingual dictionary using a statistical method, and then iteratively (1) learn orthogonal mappings W_x and W_y from the previously induced bilingual dictionary, and (2) induce bilingual dictionary from previously induced mappings until convergence. In what follows, we describe each step in details.

Train mapping Supposing that D is the previously induced bilingual dictionary, we learn mappings W_x and W_y that maximizes cosine similarity of word pairs in D

$$\left(\hat{W}_x, \hat{W}_y\right) = \underset{W_x, W_y}{\arg\max} \sum_{i,j \in D} (X_i W_x) \cdot (Y_j W_y).$$

This optimization problem has an analytical solution, and the solution can be efficiently computed using singular value decomposition [1].

Induce bilingual dictionary Using previously trained mappings, $\hat{W}_x$ and $\hat{W}_y$, we now induce an updated bilingual dictionary which will be used in the next iteration to train the mappings. From every word from both of the languages, we take the nearest neighbor in the opposite language.

Due to high dimensionality, computing nearest neighbors with cosine similarity suffers from hubness problem [7], where a few points become the nearest neighbor of many other points. To mitigate this problem, they use cross-domain similarity local scaling (CSLS) [2] instead of cosine similarity in the computation of the nearest neighbor. To compute CSLS similarity, we first compute mean similarity of the mapped embeddings $W_x X_s$ of a source word s and its neighbors in the target language as follows

$$r_T(W_x X_s) = \frac{1}{K} \sum_{t \in \mathcal{N}_T} \cos(W_x X_s, W_y Y_t)$$

where $\mathcal{N}_T$ is the set of K neighbors of s in the target language. Likewise, we also compute $r_S(W_y Y_t)$ for each word t in the target language. Now, we compute CSLS similarity of a word s in the source language and a word t in the target language by

$$\text{CSLS}(W_x X_s, W_y Y_t) = 2\cos(W_x X_s, W_y Y_t) - r_T(W_x X_s) - r_S(W_y Y_t).$$

To enhance robust learning, they apply random dropout to the similarity matrix before dictionary induction.

Chapter 4

Cross-lingual word embeddings based on subword alignment

To mitigate the problem of ambiguous translations in distant language pairs we discussed in § 1.2, we take advantage of words that have surface correspondences such as loanwords and named entities, and use these word pairs for inducing cross-lingual word embeddings. We assume that such word pairs with surface correspondences are likely to be unambiguously translatable with each other since those words are originally came from the other language.

To extract such words from the bilingual dictionary, we exploit subword alignment to extract *well-aligned* word pairs in the bilingual dictionary. We first prepare an initial bilingual dictionary by exploiting existing unsupervised method to obtain cross-lingual word embeddings [1]. We then train the subword alignment model [38] from the bilingual dictionary to assign an alignment score to each word pair in the dictionary. Word pairs with greater alignment scores are extracted to create an unambiguously translatable bilingual dictionary as they are expected to be loanwords or named entities. When combined with the unsupervised method of bilingual dictionary induction [1], our method can work in a fully unsupervised manner and does

not rely on any cross-lingual resources such as a bilingual dictionary or a parallel corpus through entire steps.

The contributions of this study are as follows:

- We experimentally confirmed that the quality of cross-lingual word embeddings obtained through an existing method is degraded in distant language pairs.
- We proposed a novel method to obtain cross-lingual word embeddings that exploit subword alignment.
- Our method sets the new state-of-the-art for the task of inducing cross-lingual word embeddings for distant language pairs without supervision.

The structure of this chapter is as followed. In § 4.1, we propose a novel method to obtain cross-lingual word embeddings by exploiting subword alignment. In § 4.2, we conduct a seriese of experiments to evaluate our method and understand its characteristics. Finally, we will summarize this work in § 4.3.

4.1 Proposal

Here, we explain the details of our method to obtain cross-lingual word embeddings of two languages by exploiting subword alignment. To address a common situation where no hand-built bilingual resource is available, we design our method to be fully unsupervised; no hand-built bilingual resource is required in any steps.

Our method first obtains an initial bilingual dictionary in an unsupervised manner by exploiting the existing unsupervised method to obtain cross-lingual word embeddings [1] (explained in § 3.2). Then, we train a subword alignment model to compute an alignment score for each word pair in the dictionary. Word pairs with high alignment scores are collected to construct a refined bilingual dictionary which

we expect to contain mostly unambiguously translatable word pairs. The refined bilingual dictionary is finally used to re-train cross-lingual word embeddings.

We hereafter explain each step of our method in detail:

Step 1: Inducing initial dictionary in an unsupervised manner The first step is to obtain an initial bilingual dictionary without relying on any cross-lingual resources. For this purpose, we first train cross-lingual word embeddings of the two languages using an existing unsupervised method as described in § 3.2. Now, let X and Y be the obtained cross-lingual word embeddings of the source and target language respectively, then we construct a bilingual dictionary by taking the nearest word in the target language for each word in the source language. We employed CSLS with the neighborhood size of 10 to mitigate the hubness problem as described in § 3.2 to compute the similarity.

Step 2: Learning subword alignment model Given the bilingual dictionary induced in the previous step, we train a subword alignment model that computes the likelihood of character-level alignment of word pairs in the dictionary. For this purpose, we exploit a many-to-many alignment method [38] that is capable of aligning two sequences of symbols (words) for any language pairs. We expect this model to learn how words are imported from one language to another.

Suppose $D_{init} = \{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}$ be the bilingual dictionary induced in the previous step, where x_i and y_i are words (sequence of characters) in the source and the target languages. For each word pair (x_i, y_i), we want to find alignment $\vec{u}$ that most likely to happen.

$$\hat{\vec{u}} = \underset{\vec{u} \in \mathcal{U}_{(x_i, y_i)}}{\arg\max} P(\vec{u} | (x_i, y_i))$$

where $\mathcal{U}_{(x_i, y_i)}$ is the set of all possible alignment of x_i and y_i. This model is trained by an Expectation-Maximization algorithm.

Step 3: Filtering the initial bilingual dictionary Now, we filter the bilingual dictionary induced in Step 1 so that we obtain word pairs that have less ambiguity in mutual translation. For each word pairs (x_i, y_i), we compute the best character-level alignment $\hat{\vec{u}}$ and its alignment score, $\log P(\hat{\vec{u}}|(x_i, y_i))$. We extract word pairs with alignment scores higher than a threshold to construct the refined bilingual dictionary $D_{refined}$.

An issue here is that we may not have any development set to turn the threshold because we do not want to rely on any hand-built bilingual resources to maximize its applicability. To find the best threshold for the alignment score, we take 100 word pairs in the induced dictionary with the highest alignment scores to be a development set which we use to evaluate the resulting cross-lingual word embeddings we obtain in the next step; we adopt the threshold that achieves the best performance on the bilingual lexicon induction task for the development set. We denote the remaining bilingual dictionary as $D'_{refined}$.

Step 4: Re-training cross-lingual word embeddings We now train cross-lingual word embeddings from the reliably subword-aligned (hopefully, unambiguously translatable) bilingual dictionary that we obtained in the previous step. We employ an existing method for supervised training of cross-lingual word embeddings [39].

Given word embeddings of the source and the target languages, X and Y, and the refined bilingual dictionary $D'_{refined}$, this method trains two mappings W_x and W_y so that the mapped embeddings XW_x and YW_y are in the same semantic space, e.g., word pairs in the bilingual dictionary $D'_{refined}$ become similar after mapping. To enhance the quality of cross-lingual word embeddings, embeddings are normalized and whitened so that different components have unit variance and be uncorrelated before learning mappings and de-whitened to restore the original variance after. Like many other methods [2, 17, 18], the mappings are constrained to be orthogonal. For the details, please refer to the original paper.

4.2 Evaluation

To examine the effect of exploiting subword alignment and gain a profound understanding of our method, we conduct experiments in two distant language pairs, English-Japanese (en-ja) and English-Finnish (en-fi) and two similar language pairs, English-Spanish (en-es) and English-Italic (en-it). Following existing studies [2, 17, 39], we used the bilingual lexicon induction task for evaluation.

4.2.1 Settings

In the following, we explain the details of the experimental settings.

Bilingual lexicon induction Bilingual lexicon induction is a task to predict the translation in the target word from a word in the source language. Given a word in the source language, we take the closest word in the target language, and if the word is in the set of translations of the source word in the ground truth bilingual dictionary, we consider it to be correct. For the ground truth bilingual dictionary, we used the test set of MUSE bilingual dictionary[1] which are used in previous studies [1, 2].

For Method #2 and #3, we kept 100 word pairs of the induced initial dictionary with highest CSLS similarities for a development set to tune the filtering threshold of CSLS similarity and alignment score, respectively, and the remaining word pairs ar e used as training set. For Method #4 through 6, we used the training set of MUSE bilingual dictionary as an annotated hand-built bilingual dictionary. For Method #5 and #6, we randomly sampled 500 word pairs from this bilingual dictionary as a development set, and the remaining word pairs are used as training set. We used the development set to tune the filtering threshold of alignment score.

[1] https://github.com/facebookresearch/MUSE

We conducted Wilcoxon singed-rank test with $p = 0.05$ to show the statistically significant results in bold in results.

Monolingual word embeddings Both our method and the baseline method construct cross-lingual word embeddings from monolingual word embeddings that are trained independently. We used monolingual word embeddings obtained by applying Subword-Information Skip-gram (SISG) [37] on Wikipedia corpus[2] except in § 4.2.4. We used pre-trained word embeddings available online[3] for all language except for Japanese. For Japanese, we used the official Implementation of SISG[4] to train word embeddings from Wikipedia dump file, because the pre-trained embeddings online are broken. For all languages, we take 200,000 most frequent words as our vocabulary.

Implementation For character-level many-to-many alignment in Step 2 of our method, we used mpaligner[5] version 0.97. To learn mapping across languages in Step 1 and Step 3, we used official implementation[6] of the original papers [1, 39] with the default hyperparameters.

4.2.2 Detailed evaluation in four language pairs

First, we evaluate our method and other methods in details in four language pairs: English-Japanese (en-ja), English-Finnish (en-fi), English-Spanish (en-es), and English-Italic (en-it).

[2] https://dumps.wikimedia.org/
[3] https://github.com/facebookresearch/fastText/blob/master/pretrained-vectors.md
[4] https://github.com/facebookresearch/fastText
[5] https://osdn.net/projects/mpaligner/
[6] https://github.com/artetxem/vecmap

Methods for comparison In order to evaluate an impact of subword alignment to filter out a bilingual dictionary used to induce cross-lingual word embeddings [39], we compare six methods, including three unsupervised methods and three supervised methods, that differ in how to prepare the bilingual dictionary for inducing cross-lingual word embeddings [39].

Method #1 (unsupervised) The method described in § 3.2 [1].

Method #2 (unsupervided with naive CSLS filtering) This method filters the dictionary finally used in Method #1 using CSLS similarities of the word pairs.

Method #3 (unsupervised with our subword alignment-based filtering) Our method described in § 4.1 filters the dictionary finally used in Method #1 by subword alignment.

Method #4 (supervised with a hand-built bilingual dictionary) This method uses a hand-built bilingual dictionary (described below) [39].

Method #5 (supervised with a refined hand-built bilingual dictionary) This method filters the hand-built bilingual dictionary by subword alignment.

Method #6 (supervised with a combined bilingual dictionary) This method combines bilingual dictionary obtained by Method #4 (with a different threshold to alignment scores) and the hand-built bilingual dictionary used in Method #3.

For Method #2 and #3, we kept 100 word pairs of the induced initial dictionary with highest CSLS similarities for a development set to tune the filtering threshold of CSLS similarity and alignment score, respectively, and the remaining word pairs ar e used as training set. For Method #4 through 6, we used the training set of MUSE bilingual dictionary as an annotated hand-built bilingual dictionary. For Method #5 and #6, we randomly sampled 500 word pairs from this bilingual dictionary as

	manual dict. (filtering)	unsup. dict. (filtering)	distant		similar	
			en-ja	en-fi	en-es	en-it
#1	-	[1]	0.4573	0.4393	0.8086	0.7713
#2	-	[1] (CSLS)	0.4440	0.4400	0.8000	0.7673
#3	-	[1] (alignment)	**0.4874**	**0.4547**	0.8087	0.7787
#4	MUSE	-	0.5175	0.4373	0.7940	0.7587
#5	MUSE (alignment)	-	0.4944	0.4320	0.7913	0.7580
#6	MUSE	[1] (alignment)	**0.5210**	**0.4766**	0.8033	0.7686

TABLE 4.1: The accuracy of bilingual lexicon induction

a development set, and the remaining word pairs are used as training set. We used the development set to tune the filtering threshold of alignment score. The results are shown in Table 4.1

Comparison with unsupervised baseline First, if we compare our Method #3 with the unsupervised baseline Method #1, our method outperforms the unsupervised baseline. By comparing the performance among different language pairs, we find that the difference is more significant for distant language pairs (en-ja, en-fi), while we gain minor improvements in similar language pairs (en-es, en-it). From these results, we confirmed the effectiveness of our method, especially for distant language pairs.

Comparison with alternative filtering method Our method filtered the induced bilingual dictionary by subword alignment to obtain an refined bilingual dictionary that consists of unambiguously translatable word pairs, and successfully obtained high-quality cross-lingual word embeddings. Here, we examine if we genuinely need subword alignment, or if other simple methods of filtering also yield similar results.

Instead of filtering the bilingual dictionary by the alignment scores (§ 4.1), we filtered it by CSLS similarity scores used in step 1 (Method #1). Like our method, 100 word pairs with highest CSLS similarity scores are kept as a development set

and are used to find the best threshold for the filtering. This filtering method is expected to yield higher quality bilingual dictionary. However, it does not consider ambiguity of words.

For all language pairs, we found that our method outperforms the alternative filtering method. Thus, we can conclude that the quality of cross-lingual word embeddings is improved only by exploiting subword alignment.

Evaluation of supervised methods Occasionally, a hand-built bilingual dictionary is available to obtain cross-lingual word embeddings. Here, we consider what method is suited in such a situation. For this purpose, we compare three methods including the supervised baseline [39] (Method #4), and two modified versions of our method (Method #5 and #6). In all of the language pairs, the highest accuracy was obtained by concatenating the hand-built bilingual dictionary with the refined dictionary that is obtained in an unsupervised manner (§ 4.1) for inducing cross-lingual word embeddings. However, if we compare supervised Method #4, #5, and #6 using a hand-crafted bilingual dictionary with the fully unsupervised Method #1, #2, and #3, the unsupervised methods outperform the supervised methods on the similar language pairs with a small margin. Method #6, the combination of the hand-crafted dictionaries and those fileted from the automatically-induced method by subword alignment yielded the best performance for the two distant-language pairs.

4.2.3 Evaluation in various language pairs

To evaluate our method in various situations, we compare our method (#3 in § 4.2.2) with unsupervised baseline method [1] (#1 in § 4.2.2) in eight additional language pairs: English-Danish (en-da), English-German (en-de), English-French (en-fr), English-Dutch (en-nl), English-Portuguese (en-pt), English-Swedish (en-sv), English-Turkish (en-tr), and English-Persian (en-fa). The result is shown in Table 4.2.

Method	en-da	en-de	en-fr	en-nl	en-pt	en-sv	en-tr	en-fa
Baseline (#1)	0.5567	0.7327	0.8040	0.7333	0.7853	0.6040	0.4827	0.3147
Ours (#3)	**0.6100**	0.7373	0.8013	0.7347	**0.8020**	**0.6233**	0.4833	0.3127

TABLE 4.2: The accuracy of bilingual lexicon induction in additional 8 language pairs

Lang.	# tweets (m)	Ave. # tokens
English	193	14.18
Japanese	117	19.32
Finnish	26	17.01
Spanish	43	14.62
Italic	93	16.47

TABLE 4.3: The statistics of Twitter corpus

In six of eight languages, our method (#3) outperformed the unsupervised baseline (#1), especially in en-da and en-pt. Furthermore, the difference in the accuracy in the other two language pairs are minimal. This result confirms the applicability of our method in various of language pairs.

4.2.4 Evaluation on Twitter corpus

The monolingual word embeddings we used in the experiments are comparable corpora rather than monolingual corpora, and it may affect the performance significantly. Therefore, we evaluated our method (#3) with the unsupervised baseline (#1) on word embeddings obtained from Twitter corpora to understand the performance in such a scenario.

We obtained raw corpora consists of tweets (excluding retweets) in 2017/8 in English, Japanese, Finnish, Spanish, and Italic. User IDs starting from "@" are replaced with a special token, and all URLs are removed. We then tokenized the corpora using NLTK[7]. We show the details of the resulting corpora in Table 4.3.

[7] https://www.nltk.org/api/nltk.tokenize.html

Method	distant		similar	
	en-ja	en-fi	en-es	en-it
Baseline (#1)	0.2898	0.7831	0.5223	0.4386
Ours (#3)	0.2810	**0.7908**	**0.5534**	**0.4428**

TABLE 4.4: The accuracy of bilingual lexicon induction on Twitter corpus

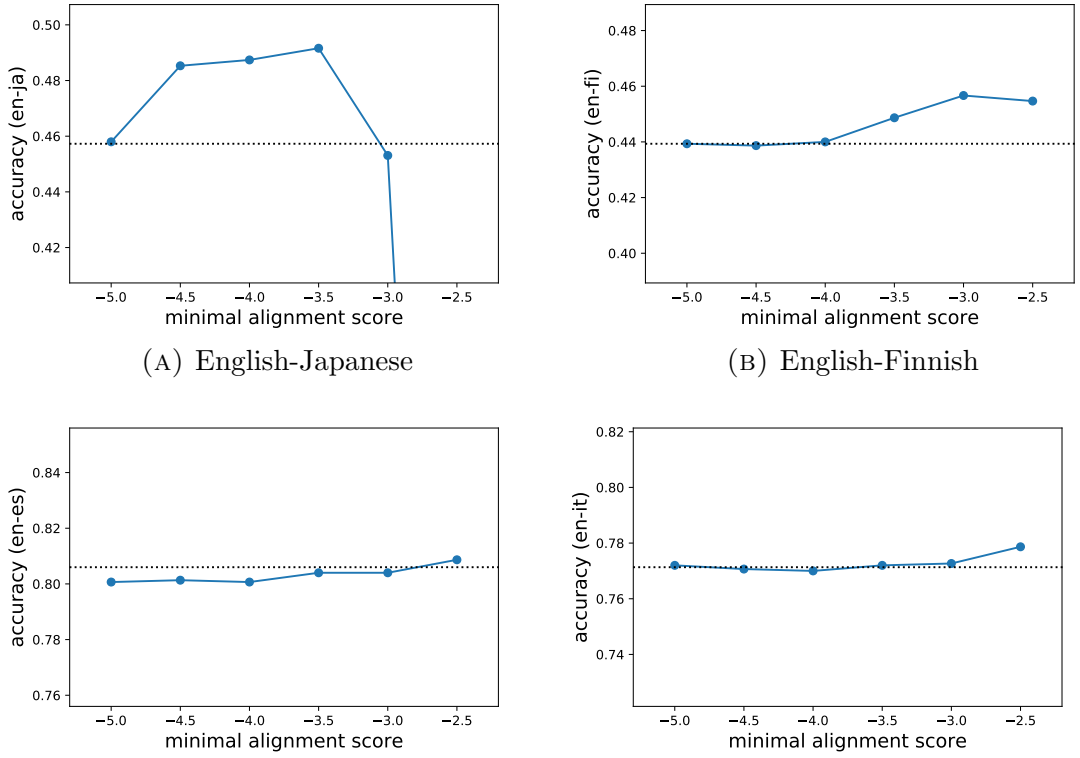
The results are shown in Table 4.4. The accuracy in English-Finnish is significantly greater than results with Wikipedia corpus shown in Table 4.1. This is because the embeddings obtained from Twitter corpora have lower coverage (35%) of the ground truth bilingual dictionary compared to embeddings obtained from Wikipedia corpora (100%). Among three of four language pairs tested, our method (#3) outperformed the unsupervised baseline method (#1), but the accuracy is generally lower than Table 4.1.

4.2.5 Qualitative analysis

Induced bilingual dictionary From the refined bilingual dictionary obtained in Step 2 (§ 4.1), we present top-10 word pairs with the highest alignment scores excluding ones with the exact same character string in Table 4.5. We also show the alignment score ranking including word pairs with exactly same character strings.

We can see that we successfully obtained loanword pairs such as cost-コスト in English-Japanese, camera-kamera in English-Finnish, and international-internacional in English-Spanish, and named entities such as india-intia in English-Finnish, and americans-americani in English-Italic.

Sensitivity of hyperparameters To find the best threshold of alignment score, we kept 100 word pairs in the induced bilingual dictionary with the highest alignment scores as development set. Here, we consider how sensitive the resulting word embeddings are to the value of hyperparameters.



(A) English-Japanese (B) English-Finnish

(C) English-Spanish (D) English-Italic

FIGURE 4.1: The accuracy of bilingual lexicon induction task with various value of threshold. The dotted line indicates the unsupervised baseline (#1).

In Table 4.1, we show the relationship between minimal alignment score and the accuracy of the bilingual lexicon induction task. For all of the language pairs except for English-Japanese, we found a tendency to improve the accuracy as the minimal alignment score increases or the size of resulting bilingual dictionary decreases. For English-Japanese, the resulting bilingual dictionary was too small to properly induce the mapping at -0.25.

4.3 Summary

In this paper, we analyzed cross-lingual word embeddings and found that their performances degrade in distant language pairs. To mitigate this problem, we proposed a novel unsupervised method to obtain cross-lingual word embeddings by exploiting subword alignment to utilize unambiguously translatable words.

In experiments, our method outperformed the state-of-the-art unsupervised and supervised method to obtain cross-lingual word embeddings, especially for distant language pairs, and advanced the new state-of-the-art for bilingual lexicon induction. Through analysis, we confirmed that our method correctly identifies loanwords and named entities that are expected to be helpful to obtain cross-lingual word embeddings as they tend to have less ambiguity.

rank	English	Japanese
1	chart	チャート *(tya a to)*
2	demonstration	デモンストレーション *(de mo n su to re e sho n)*
3	plantation	プランテーション *(pu ra n te e sho n)*
4	sparta	スパルタ *(su pa ru ta)*
5	elf	エルフ *(e ru hu)*
6	scrap	スクラップ *(su ku ra ppu)*
7	ana	アナ *(a na)*
8	timing	タイミング *(ta i mi n gu)*
9	scandal	スキャンダル *(su kya n da ru)*
10	brest	ブレスト *(bu re su to)*

(A) English-Japanese

rank	English	Finnish
68	croatia	kroatia
138	constantin	konstantin
139	israelis	israelin
196	india	intia
213	socrates	sokrates
227	camera	kamera
286	macedonian	makedonian
326	atlantic	atlantin
332	tina	nina
336	caucasian	kaukasian

(B) English-Finissh

rank	English	Spanish
323	international	internacional
487	secretaries	secretarios
496	territories	territorios
591	mercenaries	mercenarios
606	initial	inicial
628	rational	racional
653	residential	residencial
666	national	nacional
702	narrator	narrador
705	salaries	salarios

(C) English-Spanish

rank	English	Italic
439	italians	italiani
453	terrorists	terroristi
502	errors	errori
532	senators	senatori
558	arrests	arresti
616	tensions	tensioni
625	americans	americani
657	assassins	assassini
658	continents	continenti
688	aliens	alieni

(D) English-Italic

TABLE 4.5: Word pairs in the induced dictionary

Chapter 5

Fully task-specific multilingual model using cross-task projection of cross-lingual word embeddings

To exploit a fully task-specific neural network in cross-lingual settings, we propose a novel method of projecting pre-trained cross-lingual word embeddings to word embeddings of a task-specific neural network that is trained for the target task with the training data in a source language (Figure 5.1). We then utilize the obtained cross-task projection to obtain task-specific cross-lingual word embeddings of the target language that can be used for the task-specific neural network.

To obtain the above cross-task projection of cross-lingual word embeddings, we propose a simple, yet effective method of locally-linear mapping. This method is built on the assumption that local topology is preserved between the semantic spaces of word embeddings in two NLP tasks; in other words, *adequately close* words in pre-trained cross-lingual word embeddings will have similar representation even in task-specific semantic space. We first represent general cross-lingual word embedding of a word in the target language by linear combinations of general cross-lingual word embeddings of k neighboring words in the source language. We then use the

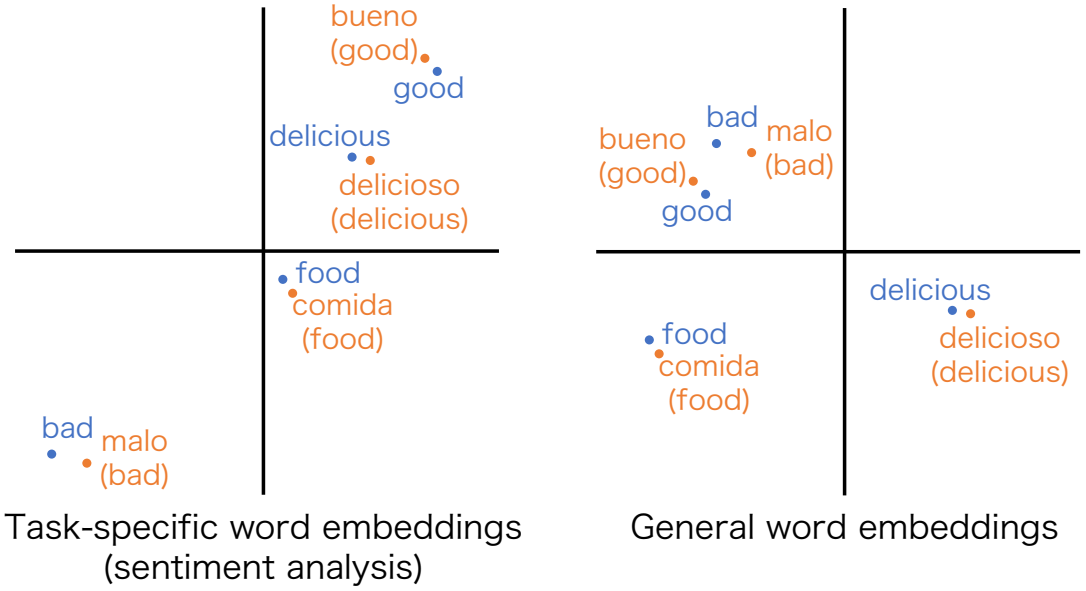


FIGURE 5.1: Conceptual diagram of task-specific cross-lingual word embeddings (right) compared to general cross-lingual word embeddings (left). *good* and *bad* are close to each other in the general semantic space while they are apart in the task-specific semantic space for sentiment analysis.

obtained weights to compute a task-specific word embedding of the target word by a linear combination of task-specific word embeddings of the k neighboring source words (§ 5.1). Note that our method does not rely on any cross-lingual resources such as bilingual dictionaries and annotated corpora in the target language, and is therefore applicable to any tasks, languages, and models with word embeddings layers.

We evaluated our method on document classification and sentiment analysis tasks. We first obtained task-specific neural networks for the two tasks using annotated corpora in the source language (English), and then induced task-specific cross-lingual word embeddings for the target languages (Spanish, German, Danish, Dutch, French, Italian, Portuguese, Swedish and Turkish) to apply the obtained neural network to those languages. Experimental results confirmed that our method successfully improved the classification accuracy of the multilingual model [14] in all

of the task-language pairs.

The contributions of this study are as follows:

- We **confirmed the limitation of the traditional multilingual model** with embedding layers fixed to pre-trained cross-lingual word embeddings.
- We established **a method of obtaining fully task-specific multilingual models** by learning a cross-task projection from general to task-specific cross-lingual word embeddings.
- Our **cross-task projection is simple and has an analytical solution** with only one hyperparameter; the solution is a global optima.

5.1 Fully task-specific multilingual model

In this section, we propose a method of projecting general cross-lingual word embeddings to semantic space of the embedding layers of purely task-specific neural networks whose all the parameters (including embeddings) are trained for the task. We assume that an annotated corpus is only available in the source language. Using the obtained task-specific cross-lingual word embeddings, we apply the fully task-specific neural networks trained with datasets in the source (resource-rich) language to the target language without assuming any cross-lingual resources or annotated corpus in the target language. Note that our method is applicable to any tasks and models as long as they have word embedding layers for input words.

5.1.1 Method overview

The entire framework of obtaining a fully task-specific multilingual model is as follows:

Step 1: inducing cross-lingual word embeddings First, we obtain general cross-lingual word embeddings $\{X^{\text{gen}}, Y^{\text{gen}}\}$ where X^{gen} and Y^{gen} are word embeddings of the source and target language i nthe save semantic sapce from a raw corpus of each language using [1] (§ 3.2). As we discussed in § 3.2, this step can be done in an unsupervised manner without using any cross-lingual resources such as a bilingual dictionary and parallel corpus.

Step 2: training a neural network in the source language Next, we train a neural network $f(\cdot; X^{\text{spec}}, \theta)$ (including an embedding layer) on annotated corpus D in the source language. The embedding layer, X^{spec}, of the resulting neural network will be considered as task-specific word embeddings of the source language. This neural network is only applicable to the source language because we do not have task-specific word embeddings in the target language in the same semantic space with X^{spec}.

Step 3: locally linear mapping Then, we induce cross-task projection ϕ that computes task-specific word embeddings of the target language Y^{spec} from pre-trained general cross-lignual word embeddings $\{X^{\text{gen}}, Y^{\text{gen}}\}$ and task-specific word embeddings of the source language X^{spec}. We will explain the details of this mapping in § 5.1.2

Step 4: applying the neural network to target language Finally, we replace embedding layer X^{spec} of the neural network obtained in Step 2 $f(\cdot; X^{\text{spec}}, \theta)$ with Y^{spec} to obtain a neural network $f(\cdot; Y^{\text{spec}}, \theta)$ which is applicable to the target language.

5.1.2 Learning cross-task projection of embeddings using locally linear mapping

Here, we will explain the detailed construction of our cross-task projection ϕ for cross-lingual word embeddings used in Step 3 in the previous section. Given general cross-lingual word embeddings $X^{\text{gen}}, Y^{\text{gen}}$ of the source and target languages, and

task-specific word embeddins X^{spec} of the source language, we need to compute task-specific word embeddings Y^{spec} of target language in the same semantic space with X^{spec}. We introduce a mapping method inspired by locally linear embeddings [40], a dimension reduction technique, assuming that the local topology among nearest neighbors will be consistent between two NLP tasks (here, language model and the target task).

The construction of this cross-task projection has two steps. First, for each word i in the target language, we take k nearest neighbors in the source language in the semantic space of the general cross-lingual word embeddings where k is a hyper-parameter. For this purpose, we used cosine similarity as the metric instead of Euclidean distance in [40].

Then we learn α_{ij} that most successfully restore Y_i^{gen} as a linear combination by optimizing

$$\alpha_{ij} = \arg\min_{\alpha_{ij}} \left| Y_i^{\text{gen}} - \sum_{j \in N_i} \alpha_{ij} X_j^{\text{gen}} \right|$$

where N_i is the set of k nearest neighbors of a taget language word i in the source language, with constraint of $\sum_j \alpha_{ij} = 1$.

The solution to this optimization problem can be analytically given by

$$\alpha_{ij} = \frac{\sum_k C_{ijk}^{-1}}{\sum_j \sum_k C_{ijk}^{-1}}$$

where

$$C_{ijk} = \left(Y_i^{\text{gen}} - X_j^{\text{gen}}\right) \cdot \left(Y_i^{\text{gen}} - X_k^{\text{gen}}\right).$$

We then compute Y_i^{spec} using the obtained weights α_{ij} by

$$Y_i^{\text{spec}} = \sum_{j \in N_i} \alpha_{ij} X_j^{\text{spec}}.$$

The resulting Y^{spec} is thereby in the same semantic space with the task-specific word embeddings of the source language, and the local topology among nearest neighbors are preserved after projection. Our locally linear mapping has only one hyperparameter k, how many neighbors in the source language we consider for the target language, and we can find the global optima by the analytical solution with simple computation.

Hyperparameter search Even though our proposing method has only one hyperparameter k, we still want to appropriately tune its value to obtain the best performance. Typically we choose k that performs best in the development dataset in the target language. However, we assume that no annotated resource is available in the target language, and thus we cannot exploit development datasets in the target language.

To address this problem, we apply our cross-task projection to the source language with various k and then choose k with the best model performance with the resulting embeddings on the development data of the source language. In experiments, we report the results with this tuning method along with another tuning method assuming a very small development data of 100 examples in the target language

5.2 Experiments

We conduct a series of experiments to evaluate our purely task-specific multilingual models obtained by the proposed cross-task projection of cross-lingual word embeddings (§ 5.1.2). Our method is language- and task-independent and applicable to various tasks where existing multilingual models can be applicable. Following existing studies on multilingual models [14, 24], we adopted classification and sentiment analysis tasks with various languages. For all of the experiments, we use English as the source language.

Language	# samples	# ave. tokens
English	673,768	237.0
Spanish	14,997	159.0
German	86,550	195.8
Danish	8,366	172.2
Frenchi	71,292	256.9
Italian	21,594	137.5
Dutch	1,690	229.0
Potuguiss	6,263	249.0
Swidish	10,383	162.3

TABLE 5.1: Statistics of RCV1/2 corpus used for document classification

5.2.1 Settings

In the following, we explain the experimental settings including details of the two target tasks and datasets we used.

Document classification is the task of predicting the topic of a given text. For this task, we use Spanish, German, Danish, French, Italian, Dutch, Portuguese, and Swedish as the target languages.

Following existing studies [14], we use RCV1/RCV2 dataset [41] for this task which contains news text in many languages with labels from 4 categories: Corporate/Industrial, Economics, Government/Social, and Markets.

For the datasets in the source language, English corpus, we randomly selected 10,000 samples for the test set, other 10,000 samples for the development set, and the rest to be the training set. For the target languages, we sample 100 samples to be development set for alternative tuning of k (§ 5.1.2) and the rest to be the test set. The summary and statistics of the datasets are given in Table 5.1.

Sentiment analysis is a task of predicting a polarity label of the writer's attitude for a given the text. We designed this task to be a binary classification of positive and negative labels. For this task, we use Spanish, Dutch, and Turkish as the target languages.

Corpus	Lang.	# samples	# ave. tokens
Yelp Review dataset	EN	4,406,965	133.0
ABSA dataset	EN	1,513	14.0
	ES	1,411	15.1
	NL	1,148	14.1
	TR	878	9.7

TABLE 5.2: Statistics of datasets used for sentiment analysis task

To train the model in English, we use Yelp Review dataset[1] which is a set of restaurant reviews with numerical ratings in the range of 1-5 given by the reviewers. We labeled the reviews with ratings of 1 or 2 to be negative, ones with ratings of 4 or 5 to be positive, and ones with ratings of 3 are excluded. In order to balance the number of positive samples and negative samples, we downsampled positive samples as in [24]. Then, we randomly sample 100,000 sample to be the development set, another 100,000 sample to be the test set, and the remaining 4,206,965 samples to be the training set.

For evaluation in the target languages, we use ABSA dataset [42] which consists of restaurant reviews in various languages including English, Spanish, Dutch, and Turkish with annotation of polarity label of positive or negative to each sentence. For each language, we randomly sample 100 sentences to be development set for alternative tuning of k (§ 5.1.2) and rest to be the test set. The summary and statistics of the datasets are given in Table 5.2.

Preprocessing We apply the same preprocessing to all dataset we use. All corpora are tokenized by NLTK[2] tokenizer and lowercased to match vocabularies of pre-trained word embeddings.

[1] https://www.yelp.com/dataset
[2] https://www.nltk.org/api/nltk.tokenize.html

General cross-lingual word embeddings General cross-lingual word embeddings were obtained using the state-of-the-art unsupervised method with self-learning framework[3] [1] as described in § 3.2. This method takes monolingual word embeddings of two languages and learns the mapping between them to obtain cross-lingual word embeddings. For monolingual word embeddings, we used pre-trained word embeddings available online.[4] They are word embeddings with the dimension of 300 obtained by applying subword-information skip-gram [37], which is a widely used method for monolingual word embeddings, to Wikipedia corpus.

Models Our method is applicable to any neural networks with word embeddings layer as the existing multilingual models are. In the experiments, we implement a simple bag-of-embeddings model that takes the dimension-wise average of all embeddings of input tokens. The resulting vector is then fed to following single-layer feedforward neural network.

In order to evaluate the impact of task-specific word embeddings and effectiveness of our cross-task projection of cross-lingual word embeddings, we compare three different multilingual models.

GenEmb [14] trains the feed-forward neural network with embedding layers fixed to the pre-trained cross-lingual word embedding.

CrossTaskProj trains the feed-forward neural network in the target language and make it cross-lingual by cross-task projection as described in § 5.1.

EmbFFNN adds an embedding-wise two-layer feedforward neural network to the feed-forward network used in TaskSpekEmb. The additional element-wise feedforward neural network is (with embedding layers fixed to pre-trained cross-lingual word embedding) intended to map pre-trained word embeddings to task-specific word embeddings.

[3] https://github.com/artetxem/vecmap

[4] https://github.com/facebookresearch/fastText/blob/master/pretrained-vectors.md

Method	Doc. class.	Sent. analysis
GenEmb	0.842	0.809
EmbFFNN	0.958	0.853
CrossTaskProj	**0.975**	**0.870**

TABLE 5.3: Classification accuracy of models evaluated in the source language (English) on document classification and sentiment analysis

As stated in § 5.1.2, we used two strategies to tune the additional hyperparameter k for **CrossTaskProj**: 1) utilize the development set of the source language as described in § 5.1.2, and 2) prepare a very small (100 samples) development data in the target language. For document classification task, we also evaluate **CrossTaskProj** with subword alignment based cross-lingual word embeddings (SACLWE) we discussed in § 4. We run all experiments three times and report average classification accuracy.

5.2.2 Results

Here, we will report the result of our experiments and evaluate the effect of the fully optimized multilingual model. First we evaluate **GenEmb** and **CrossTaskProj** in English to understand the impact of task-specific word representation in neural networks. Then, we evaluate the models in cross-lingual settings to evaluate how well our locally linear mapping produce task-specific cross-lingual word embeddings.

Impact of task-specific word embeddings We examine the impact of optimizing word embeddings to the given task on model accuracy through experiments in English. Table 5.3 shows the result of document classification and sentiment analysis tasks respectively.

In both tasks, **CrossTaskProj** which has task-specific word embeddings outperformed **GenEmb** with a wide margin. So, task-specific word embeddings are crucial

Method	en-es	en-de	en-da	en-fr	en-it	en-nl	en-pt	en-sv
GenEmb	0.376	0.767	0.635	0.665	0.542	0.662	0.477	0.803
EmbFFNN	0.669	0.697	0.571	0.778	0.569	0.744	0.424	0.466
CrossTaskProj								
Tuned on the src. lang.	0.666	0.753	0.697	**0.854**	0.569	0.799	0.557	0.812
Tuned on the trg. lang.	**0.724**	**0.788**	**0.718**	0.840	**0.617**	**0.823**	**0.588**	**0.820**
CrossTaskProj on SACLWE								
Tuned on the src. lang.	0.488	0.693	0.672	0.748	0.492	0.816	0.514	0.800
Tuned on the trg. lang.	0.682	**0.794**	**0.738**	**0.859**	**0.621**	0.815	0.542	**0.844**

TABLE 5.4: Accuracy in document classificaiton task. All models are train on English dataset and applied in the other languages.

Method	en-es	en-nl	en-tr
GenEmb	0.802	0.736	0.695
EmbFFNN	0.773	0.705	0.679
CrossTaskProj			
Tuned on the source lang.	0.825	0.759	**0.712**
Tuned on the target lang.	**0.826**	**0.763**	0.709

TABLE 5.5: Accuracy in sentiment analysis task. All models are train on Yelp Review dataset in English and applied in the other languages. Evaluation datasets are ABSA dataset for each language.

to obtain better model performance. This result motivates us to learn task-specific cross-lingual word embeddings to exploit fully task-specific neural network.

When compared **CrossTaskProj** and **EmbFFNN**, we found that in both document classification task and sentiment analysis task **CrossTaskProj** outperforms **EmbFFNN** in monolingual evaluation. This indicates that inducing task-specific word embeddings has a more significant contribution to the model performance than having deeper models.

Performance of multilingual models We evaluate the performance of the models in cross-lingual settings. In Table 5.4 and Table 5.5, we report results of cross-lingual evaluation on sentiment analysis task and document classification task. All

models are trained in English and evaluated in other languages.

CrossTaskProj with hyperparameter tuning on the development set of the source language, English, successfully outperformed the baseline, **GenEmb**, in all language pairs. This result indicates the importance of task-specific word representation in the multilingual model and that locally linear mapping successfully induces task-specific cross-lingual word embeddings. While we gained a little more improvements by tuning k on the small (100 samples) development set in the target language, the gain is relatively small compared to the gain between **GenEmb** and **CrossTaskProj** with hyperparameter tuning on the source language development set. We gained a little more improvements by tuning k on the small (100 samples) development set in the target language.

Comparing **CrossTaskProj** and **EmbFFNN**, the difference in classification accuray in cross-lingual setting is more significant than monolingual setting. In several languages, **EmbFFNN** has even lower classification accuracy compared to **GenEmb**. We guess that by having more layers, the model becomes more sensitive to the small difference in word representation, and thus, the noise in pre-trained cross-lingual word embeddings degrades the model performance.

Performance with subword alignment based cross-lingual word embeddings A fully task-specific multilingual model relies on cross-lingual word embeddings as described in § 5.1 and we adopted an existing method [1] to obtain them. Here, we adopt subword alignment based cross-lingual word embeddings we discussed in § 4 instead, and evaluate the performance.

When the hyperparameter is tuned on the small development set, we found fully task-specific multilingual models with subword based cross-lingual word embeddings outperforms ones with embeddings obtained by the existing method in six of eight language pairs we tested. This result confirms the quality of subword alignment based cross-lingual word embeddings.

However, we observed the degradation of the performance when the hyperparameter is tuned on the development set of the source language. We guess that the subword alignment based cross-lingual word embeddings have a small bias as the bilingual dictionary is filtered and thus it is important to correctly tune the hyperparameter. This is not a significant problem because only very small supervision (100 samples) in the target language is required.

5.2.3 Analysis

Here, we further investigate the characteristics of our method. First, we visualize the task-specific cross-lingual word embeddings we obtained by a locally linear mapping for the sentiment analysis task using t-SNE [43]. We then discuss about the sensitivity of our method to hyperparameter k to obtain some valuable insights.

Task-specific sementic space of sentiment analysis In Figure 5.2, we show the visualization of general cross-lingual word embeddings and task-specific cross-lingual word embeddings for sentiment analysis task in English and Spanish. We used multilingual sentiment dataset[5] [44], which contains the list of positive words (shown as green dots) and negative words (shown as red crosses) for English and Spanish, to indicate the distribution of positive and negative words in the visualization. Also, among positive and negative words, we present five most frequent words in the ABSA test dataset.

For English, we found that task-specific word embeddings extracted from the embedding layer of trained neural network successfully learns task-specific features, and positive words and negative words are distinguishable. However, they are mixed and not distinguishable in general word embeddings. Notice that *delicious* and *best* are close in task-specific word embeddings while they are apart in general word embeddings.

[5] https://sites.google.com/site/datascienceslab/projects/multilingualsentiment

Visualization of Spanish word embeddings exhibits similar property. Positive words and negative words are distinguishable in task-specific word embeddings obtained by our cross-task projection, but they are indistinguishable in general word embeddings. This indicates that locally linear mapping successfully transfers task-specific feature of words across languages.

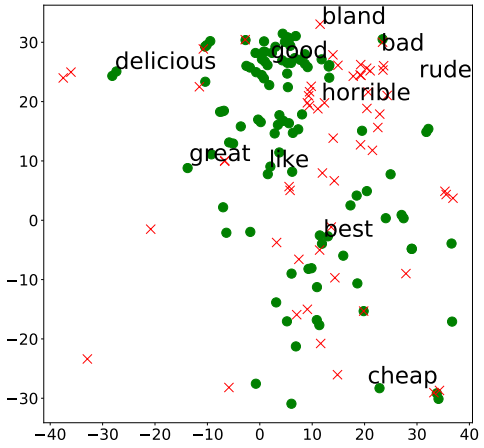
Sensitivity to hyperparameter k We proposed two strategies to tune the hyperparameter k of our cross-task projection of cross-lingual word embeddings: tuning on the development set in the source language as described in § 5.1.2 or preparing small (100 samples) development sets in the target languages. In many cases, hyperparameter search on the development set in the target languages improves the model quality even when only small development set is available. We believe this is because the best value of k is language dependent. In what follows, we evaluate how the model quality is affected by the value of k among different languages and tasks.

In Figure 5.3, we present classification accuracy of the model with various value of k in document classification task and sentiment analysis task. In many cases, $k = 1$ performs best, but in several occasion, larger value achieved better classification accuracy. For document classification task, performance of the model varies significantly by changing k for many languages, but for sentiment classification task, the performance is consistant with various value of k.

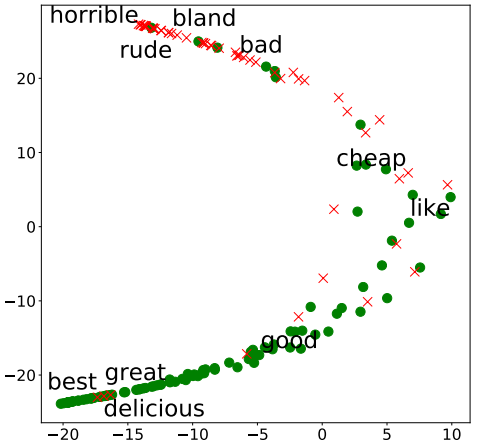
5.3 Summary

In this paper, we proposed a novel method to obtain a fully task-specific multilingual model without relying on any cross-lingual resources or annotated corpus in the target language. Our method induces task-specific cross-lingual word embeddings for the target language using our novel method of locally linear mapping.

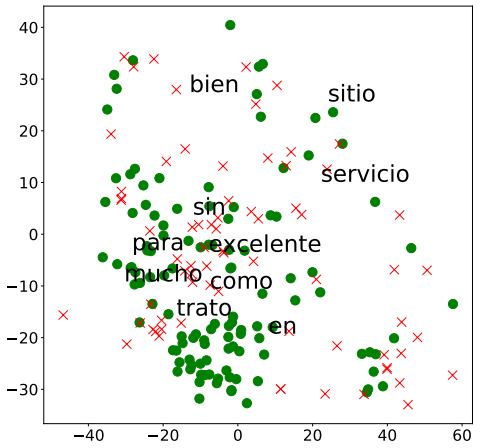
Through experiments, we showed that the true potential of a neural network is not exerted by the existing multilingual model as they fix the embedding layer of the neural network to pre-trained cross-lingual word embeddings. Experimental results and analysis confirmed that our cross-task projection successfully obtains task-specific word embeddings of the target language without any annotated resources in the target language, and classification accuracy of the resulting purely task-specific multilingual model outperformed existing multilingual model with embedding layer fixed to general word embedding with a wide margin.



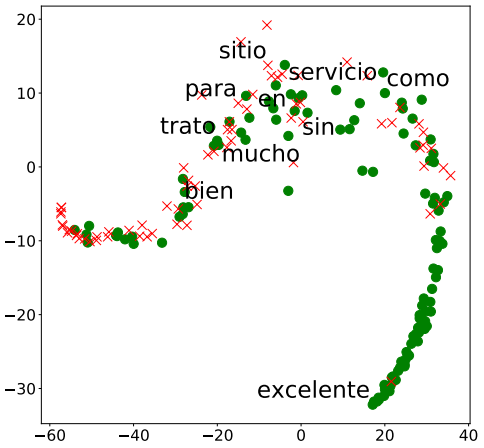
(A) General word embeddings of English



(B) Task-specific word embedding of English

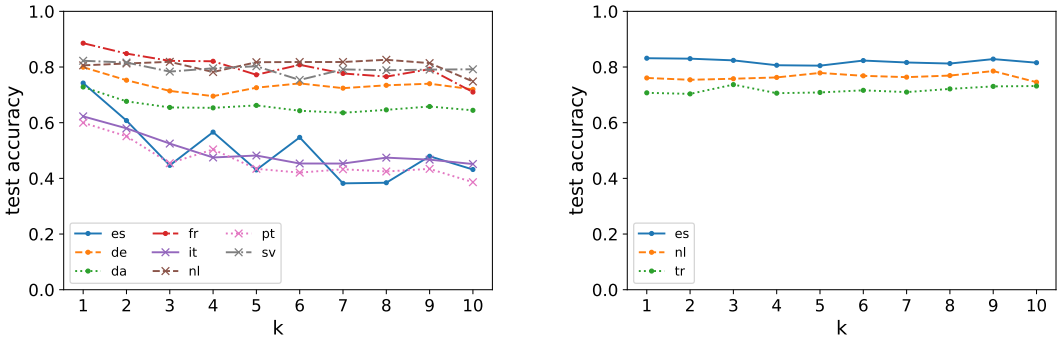


(C) Task-specific word embedding of Spanish



(D) Task-specific word embedding of Spanish

FIGURE 5.2: The t-SNE visualization of English and Spanish word embeddings in sentiment analysis



(A) Document classification (B) Sentiment analysis

FIGURE 5.3: The classification accuracy with various value of k

Chapter 6

Conclusion

In this study, we proposed two methods to improve a multilingual model. The multilingual models obtained using our method are robust to the distant languages due to cross-lingual word embeddings based on subword alignment (§ 4) and achieves better performance due to task-specific cross-lingual word embeddings (§ 5). Here, we summarize our contributions in this thesis (§ 6.1), and then discuss the future of multilingual models (§ 6.2).

6.1 Contribution of this thesis

In § 4, we first observed that the quality of cross-lingual word embeddings obtained in an unsupervised manner degrades for distant language pairs. Our method of subword alignment based cross-lingual word embeddings mitigates this problem using subword alignment for filtering the unsupervisedly induced bilingual dictionary to obtain unambiguously translatable word pairs. Empirical results (§ 4.2) confirmed that the performance in distant languages improves with our method.

In § 5, we proposed a method to obtain task-specific cross-lingual word embeddings in order to exert the true potential of neural networks by inducing task-specific word

representation. Because the most commonly used method to obtain a multilingual model fixes its embedding layer to pre-trained cross-lingual word embeddings optimized to language model task, it fails to induce task-specific representation of words. Our method mitigates this problem by learning cross-task projection, namely locally linear mapping, to map the pre-trained cross-lingual word embedding to task-specific cross-lingual word embeddings. This method is applicable to any neural network model with an embedding layer for any task. In experiments (§ 5.2), we observed improvements in various languages in two distinct tasks.

6.2 Future of multilingual model

While our method obtains fully task-specific multilingual models which is applicable to distant languages, there are many remaining issues to overcome to obtain multilingual models that perform equally well for all languages. We discuss what kind of improvements we must make to accomplish such a goal.

Further improvements of cross-lingual word embeddings in distant language pairs As discussed in § 4, we successfully improved the quality of cross-lingual word embeddings in distant language pairs. However, the accuracies in bilingual lexicon induction in distant language pairs are still not comparable to ones in similar language pairs. We believe this is due to the difference in grammar (word order) and word segmentation across languages. In most of existing methods, including ours, the translation of a word is assumed to be a single word, but in reality, a word often translated into multiple words especially in distant language pairs. For example, the translation of an English word "she" in Japanese is "彼女の" which consists of two words. We guess that this difference is the crucial problem that degrades the quality of cross-lingual word embeddings in distant language pairs.

Absorbing gramatical difference among languages In this study, we consider a multilingual model that utilize cross-lingual word embeddings to absorb the difference in vocabularies across languages. However, such methods fail to absorb the difference in grammar or word order. Therefore, we believe that our method and any other methods that simply utilize cross-lingual word embeddings fails to work well with neural networks that incoporate sequential information such as Convolutional Neural Network [45] and Long Short-Term Memory [46] which are known to work well in many NLP tasks.

For this purpose, several methods are available such as pre-training a multilingual encoder from raw corpora and incorporating a machine translation model to absorb the difference in grammar. A pre-trained multilingual encoder takes a sentence (or sequence of sentences) and produces language-independent representation of the input, and pre-trained cross-lingual word embeddings can be considered as a kind of pre-trained multilingual encoder. Pre-trained cross-lingual word embeddings can be considered a kind of pre-trained multilingual encoder, and thus pre-trained multilingual encoder suffers from the same issue with pre-trained cross-lingual word embeddings we discussed in § 1.3. Thus, the multilingual encoder has to have enough representative power so that it is applicable to various tasks, or we must have some method to fine-tune the encoder while keeping it multilingual.

In several existing studies, a machine translation model was used to obtain a cross-lingual model [24, 25]. A problem regarding these methods is that it requires massive parallel corpus to obtain high-quality machine translation model, and the noise in the translation affects the model performance significantly. However, in recent years, several unsupervised methods to train machine translation models without any cross-lingual corpus are proposed [47–53], and the performance is increasing. Thus, we believe it worth investigating the possibility of exploiting unsupervised machine translation for multilingual models.

Exploiting different types of resources In this study, we assumed that there is an annotated corpus in the source language and there is neither an annotated

corpus in the target language nor cross-lingual resources across the source and the target languages. Furthermore, we did not consider any other languages except these two languages. However, in reality, there are many other languages which may or may not have annotated corpora, and for some language pairs, we may obtain cross-lingual resources. To accomplish the best performance, it is important to exploit all of these resources effectively.

However, our methods and many other methods of obtaining a multilingual model and cross-lingual word embeddings fail to do so. It is our future work to solve this issue to enhance the quality of multilingual models.

Bibliography

[1] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018.

[2] Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. In *Proceedings of the 6th International Conference on Learning Representation*, 2018.

[3] Yoon Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, 2014.

[4] B. Shi, Z. Fu, L. Bing, and W. Lam. Learning domain-sensitive and sentiment-aware word embeddings. In *ACL*, 2018.

[5] Tomas Mikolov, Quoc V Le, and Ilya Sutskever. Exploiting Similarities among Languages for Machine Translation. *arXiv:1309.4168*, 2013.

[6] Yutaro Shigeto, Ikumi Suzuki, Kazuo Hara, Masashi Shimbo, and Yuji Matsumoto. Ridge regression, hubness, and zero-shot learning. In *Proceedings of the 2015th European Conference on Machine Learning and Knowledge Discovery in Databases - Volume Part I*, ECMLPKDD'15, pages 135–151, Switzerland, 2015. Springer.

[7] Georgiana Dinu and Marco Baroni. Improving zero-shot learning by mitigating the hubness problem. In *Proceedings of the 3rd International Conference on Learning Representation (ICLR 2015)*, 2015.

[8] Manaal Faruqui and Chris Dyer. Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, 2014.

[9] Ang Lu, Weiran Wang, Mohit Bansal, Kevin Gimpel, and Karen Livescu. Deep multilingual correlation for improved word embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 250–256. Association for Computational Linguistics, 2015.

[10] Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2015.

[11] Yuan Zhang, David Gaddy, Regina Barzilay, and Tommi Jaakkola. Ten pairs to tag – multilingual pos tagging via coarse mapping between embeddings. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1307–1317. Association for Computational Linguistics, 2016.

[12] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2289–2294. Association for Computational Linguistics, 2016.

[13] Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A. Smith. Massively multilingual word embeddings. *arXiv:1602.01925*, 2016.

[14] Long Duong, Hiroshi Kanayama, Tengfei Ma, Steven Bird, and Trevor Cohn. Multilingual training of crosslingual word embeddings. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 2017.

[15] Meng Zhang, Yang Liu, Huanbo Luan, Yiqun Liu, and Maosong Sun. Inducing bilingual lexica from non-parallel data with earth mover's distance regularization. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. The COLING 2016 Organizing Committee, 2016.

[16] Ndapa Nakashole. NORMA: Neighborhood sensitive maps for multilingual word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2018.

[17] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, 2017.

[18] Samuel L. Smith, David H. P. Turban, Steven Hamblin, and Nils Y. Hammerla. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *Proceedings of the Fifth International Conference on Learning Representations (ICLR)*, 2017.

[19] Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. Earth mover's distance minimization for unsupervised bilingual lexicon induction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1934–1945. Association for Computational Linguistics, 2017.

[20] Martin Arjovsky and Leon Bottou. Towards principled methods for training generative adversarial networks. In *Proceedings of the 5th International Conference on Learning Representations (ICLR)*.

[21] Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. Adversarial training for unsupervised bilingual lexicon induction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1959–1970. Association for Computational Linguistics, 2017.

[22] Takashi Wada and Tomoharu Iwata. Unsupervised cross-lingual word embedding by multilingual neural language models. *arXiv:1809.02306 [cs]*, 2018.

[23] Stephan Gouws and Anders Søgaard. Simple task-specific bilingual word embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2015.

[24] Xiaojun Wan. Co-training for cross-lingual sentiment classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 235–243. Association for Computational Linguistics, 2009.

[25] Akiko Eriguchi, Melvin Johnson, Orhan Firat, Hideto Kazawa, and Wolfgang Macherey. Zero-shot cross-lingual classification using multilingual neural machine translation. *arXiv:1809.04686 [cs]*, 2018.

[26] Melvin Johnson, Mike Schuster, Quoc Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernand a Vi Ã© gas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 2017.

[27] Peter Prettenhofer and Benno Stein. Cross-Language Text Classification Using Structural Correspondence Learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1118–1127. Association for Computational Linguistics, 2010.

[28] Kui Xu and Xiaojun Wan. Towards a universal sentiment classifier in multiple languages. In *Proceedings of the 2017 Conference on Empirical Methods in*

Natural Language Processing, pages 511–520. Association for Computational Linguistics, 2017.

[29] Nikolaos Pappas and Andrei Popescu-Belis. Multilingual hierarchical attention networks for document classification. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2017.

[30] Shyam Upadhyay, Manaal Faruqui, Gokhan Tur, Dilek Hakkani-Tur, and Larry Heck. (almost) zero-shot cross-lingual spoken language understanding. In *Proceedings of the IEEE ICASSP*, 2018.

[31] Xilun Chen, Yu Sun, and Athiwaratkun Ben. Adversarial deep avaraging networks for cross-lingual sentiment classification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 46–59, 2017.

[32] Xilun Chen, Yu Sun, Ben Athiwaratkun, Claire Cardie, and Kilian Weinberger. Adversarial deep averaging networks for cross-lingual sentiment classification. *Transactions of the Association for Computational Linguistics*, 2018.

[33] Joo-Kyung Kim, Young-Bum Kim, Ruhi Sarikaya, and Eric Fosler-Lussier. Cross-lingual transfer learning for pos tagging without cross-lingual resources. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2832–2838. Association for Computational Linguistics, 2017.

[34] Zhilin Yang, Ruslan Salakhutdinov, and William W. Cohen. Transfer learning for sequence tagging with hierarchical recurrent networks. 2017.

[35] Zelling Harris. Distributional structure. In *Word*, 1954.

[36] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *Proceedings of the First International Conference on Learning Representations (ICLR)*, 2013.

[37] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 2017.

[38] Keigo Kubo, Hiromichi Kawanami, Hiroshi Saruwatari, and Kiyohiro Shikano. Unconstrained manyto-many alignment for automatic pronunciation annotation. In *Proceedings of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference 2011 (APSIPA2011)*, 2011.

[39] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *AAAI*, 2018.

[40] Sam T. Roweis and Lawrence K. Saul. Nonlinear dimensionality reduction by locally linear embedding. 290(5500):2323–2326, 2000.

[41] David D. Lewis, Yiming Yang, Tony G. Rose, and Fei Li. Rcv1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5:361–397, 2004.

[42] Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphee De Clercq, Veronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Núria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. Semeval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30. Association for Computational Linguistics, 2016.

[43] L.J.P.V.D. Maaten and GE Hinton. Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 01 2008.

[44] Yanqing Chen and Steven Skiena. Building sentiment lexicons for all major languages. In *Proceedings of the 52nd Annual Meeting of the Association for*

Computational Linguistics (Volume 2: Short Papers), pages 383–389. Association for Computational Linguistics, 2014.

[45] Yann Lecun, L é on Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, pages 2278–2324, 1998.

[46] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 1997.

[47] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Unsupervised statistical machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2018.

[48] Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. Unsupervised neural machine translation. 2018.

[49] Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Unsupervised machine translation using monolingual corpora only. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2018.

[50] Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2018.

[51] Benjamin Marie and Atsushi Fujita. Unsupervised neural machine translation initialized by unsupervised statistical machine translation. 2019.

[52] Zhen Yang, Wei Chen, Feng Wang, and Bo Xu. Unsupervised neural machine translation with weight sharing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2018.

[53] Guillaume Lample and Alexis Conneau. Cross-lingual language model pretraining. 2019.

Publications

Publications related to the thesis

International conferences

- Jin Sakuma and Naoki Yoshinaga. "Fully task-specific multilingual model using cross-task projection of cross-lingual word embeddings." Submitted to *NAACL*, 2019.
- Jin Sakuma and Naoki Yoshinaga. "Unsupervised Cross-lingual Word Embeddings Based on Subword Alignment." Submitted to *CICLing*, 2019.

Domestic conferences

- 佐久間仁，吉永直樹．"単語分散表現のタスク横断写像に基づく高精度多言語モデル．" 言語処理学会年次大会，2019

Publications non-related to the thesis

International conferences

- Masato Neishi*, Jin Sakuma*, Satoshi Tohda*, Shonosuke Ishiwatari, Naoki Yoshinaga, and Masashi Toyoda. "A Bag of Useful Tricks for Practical Neural Machine Translation: Embedding Layer Initialization and Large Batch Size." In *Proc. WAT*, AFNLP, 2017. (* Joint First Author)

Domestic conferences

- 佐久間仁，吉永直樹．"表層類似性を用いた多言語単語分散表現の教師なし学習手法." 第 233 回 自然言語処理研究会，2017.
- 根石将人，佐久間仁，遠田哲史，石渡祥之佑，吉永直樹，豊田正史．"ニューラル機械翻訳における埋め込み層の教師なし事前学習" 第 233 回 自然言語処理研究会，2017.